СЕКЦИЯ «ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ И ТЕХНИЧЕСКАЯ КИБЕРНЕТИКА»

СТАТИСТИЧЕСКИЙ АНАЛИЗ ТЕКСТОВОЙ ИНФОРМАЦИИ В СОЦИАЛЬНЫХ СЕТЯХ

Д.С. Баканов, А.В. Куприянов

Самарский национальный исследовательский университет имени академика С.П. Королева, Самара, Россия

Обоснование. В последнее время все большую роль в повседневной жизни современного человека стали играть социальные сети. Социальные сети или социальные медиа — это интернет-ресурс, который предназначен для взаимодействия людей в группах, развлечений и прочей активностей. Самая главная особенность данного ресурса состоит в том, что контент создается самими пользователями — реальными людьми и организациями. Социальные сети постепенно охватывают все большую аудиторию. Так, за период с февраля по март 2022 г. дневная аудитория ВКонтакте в России выросла на 4 млн человек, а за день ею пользуются 50 млн человек [1]. Поэтому такая платформа может послужить хорошим местом для проведения социальных исследований. Такие исследования можно использовать в различных сферах нашей жизни для улучшения качества услуг или рекламы.

Цель — произвести статистический анализ текстовой информации из социальной сети и разработать соответствующую статистическую модель.

Методы. Весь процесс анализа социальных сетей можно свести к следующим этапам:

1. Аутентификация. Пользователь при помощи открытого протокола авторизации (OAuth) входит в веб-приложение по определенному адресу и попадает в среду социальной сети.
2. Сбор данных. Данный этап зависит от особенностей социальной сети: наличие/отсутствие API, политика конфиденциальности и пр.
3. Очистка и предобработка данных.
4. Построение модели и анализ.
5. Представление результатов [2].

В данной работе рассматривается контент из группы социальной сети ВКонтакте. При помощи официального API ВКонтакте [3] было скачано свыше 80 000 постов. Их поверхностный анализ приведен в таблице.

Таблица. Результаты поверхностного анализа текстов из постов

Показатель	Значения
Всего постов	87994
Общие количественные характеристики элементов текста	Всего слов: 6087078 Всего глаголов: 1140503 Всего существительных: 1503712 Количество слов, обозначающих конкретных персон: 0
Наиболее частые представители различных элементов текста	Среди всех слов: и, в, не, я Существительные: лет, раз, мама, день Глаголы: нет, есть, могу, говорит

Для дальнейшего анализа текст был разбит на *N*-граммы — последовательность из *N* элементов строки. В данной работе — на биграммы (последовательность из двух элементов). Были найдены самые частые в использовании биграммы (рис. 1).

Из поверхностного анализа и частым биграммам можно сделать вывод, что специфика данной группы заключается в том, что в ней люди публикуют обезличенные истории из своей жизни.

На следующем этапе был проведен кластерный анализ, который заключается в сегментировании данных на кластеры (подмножества), что объекты внутри более тесно связаны, чем с другими [4].

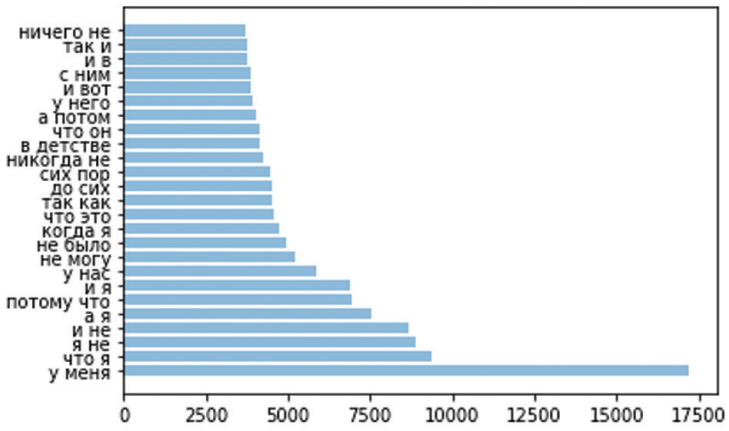

Рис. 1. Самые часто встречаемые биграммы

На следующем этапе был создан конвейер для построения статистической модели:

1. Токенизация. Текст разбивается на биграммы.
2. Векторизация. Создаются векторы пар (биграмма, количество вхождений в текст).
3. Нормализация. Векторы нормализуются в соответствии с метрикой TF-IDF. Суть данной метрики заключается в следующем: если слово встречается часто в документе и редко во всем наборе, вероятно, оно является очень представительным для этого конкретного документа и поэтому заслуживает более высокого веса [2].
4. Алгоритм *k*-средних. Данный алгоритм ищет центры кластеров, основываясь на минимизации евклидова расстояния между объектами кластера [4].

Результаты. После построения конвейера данные были разбиты на три кластера (рис. 2).

В ходе более детального исследования выяснилось, что посты были сгруппированы в основном по эмоциональному окрасу: негативный, положительный и нейтральный.

	bigram_cluster_1	bigram_freq_cluster_1	bigram_cluster_2	bigram_freq_cluster_2	bigram_cluster_3	bigram_freq_cluster_3
0	что это	1305	что это	2196	что это	1065
1	то что	900	то что	1401	то что	841
2	том что	767	том что	1184	том что	608
3	говорит что	568	день рождения	662	говорит что	511
4	думала что	408	думала что	620	день рождения	322
5	это время	401	а то	594	что её	318
6	молодой человек	392	это время	565	это время	277
7	друг друга	348	всё это	555	думала что	264
8	день рождения	316	следующий день	528	следующий день	260
9	следующий день	305	всё равно	506	что мама	230
10	всё это	274	друг друга	501	всё это	224
11	молодым человеком	268	самом деле	454	тем что	222
12	всё равно	248	тем что	411	всё равно	205
13	тем что	210	говорит что	408	друг друга	200
14	друг другу	193	както раз	400	както раз	193
15	както раз	189	молодой человек	368	а мама	170
16	а то	181	пару дней	320	а то	168

Рис. 2. Пример разбиения биграмм на кластеры

Выводы. В ходе данной работы была проанализирована текстовая информация постов группы из социальной сети ВКонтакте, построена статистическая модель кластеризации с использованием алгоритма *k*-средних.

Ключевые слова: анализ социальных сетей; кластерный анализ; алгоритм *k*-средних; наука о данных; обработка естественного языка; *N*-грамма; TF-IDF.

Список литературы

1. vk.com [Электронный ресурс]. Дневная аудитория ВКонтакте выросла на 4 млн — до рекордных 50 млн пользователей // Новости ВКонтакте [дата обращения: 01.04.2022]. Доступ по ссылке: https://vk.com/press/users-monthly-activity#:~:text=14%20марта%202022%20ВКонтакте.%20Дневная,приводят%20на%20платформу%20своих%20знакомых
2. Бонцанини М. Анализ социальных медиа на Python / пер. с анг. А.В. Логунова. Москва: ДМК Пресс, 2018. 288 с
3. dev.vk.com [Электронный ресурс]. Использование API. Быстрый старт // VK для разработчиков [дата обращения 24.02.2022]. Доступ по ссылке: https://dev.vk.com/api/getting-started
4. Хасти Т., Тибширани Р., Фридман Д. Основы статистического обучения: интеллектуальный анализ данных, логический вывод и прогнозирование, 2-е изд. / пер. с англ. Санкт-Петербург: ООО “Диалектика”, 2020. 768 с.

Сведения об авторах:

Дмитрий Сергеевич Баканов — студент, группа 6132-010402D, институт информатики и кибернетики; Самарский национальный исследовательский университет им. С.П. Королева, Самара, Россия. E-mail: dima.bakanov.1999@mail.ru

Александр Викторович Куприянов — научный руководитель, доктор технических наук, доцент; заведующий кафедрой технической кибернетики; Самарский национальный исследовательский университет им. С.П. Королева, Самара, Россия. E-mail: akupr@ssau.ru